The Art of Workload Selection

5-1

Overview

- Services Exercised
  - Example: Timesharing Systems
  - Example: Networks
  - Example: Magnetic Tape Backup System
- Level of Detail
- Representativeness
- Timeliness
- Other Considerations in Workload Selection

5-2

The Art of Workload Selection

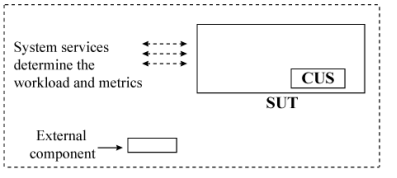
Considerations:

- Services exercised
- Level of detail
- Loading level
- Impact of other components
- Timeliness

5-3

Services Exercised

- SUT = System Under Test
- CUS = Component Under Study



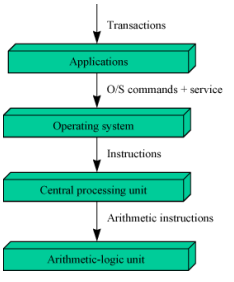
5-4

Services Exercised (Cont)

- Do not confuse SUT w CUS
- Metrics depend upon SUT: MIPS is ok for two CPUs but not for two timesharing systems.
- Workload: depends upon the system.
- Examples:
  - CPU: instructions
  - System: Transactions
  - Transactions not good for CPU and vice versa
  - Two systems identical except for CPU
    - Comparing Systems: Use transactions
    - Comparing CPUs: Use instructions
  - Multiple services: Exercise as complete a set of services as possible.

5-5

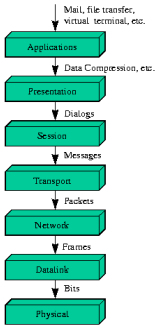
Example: Timesharing Systems



- Applications
) Application benchmark
- Operating System
) Synthetic Program
- Central Processing Unit
) Instruction Mixes
- Arithmetic Logical Unit
) Addition instruction

5-6

Example: Networks



Level of Detail

- Most frequent request:
  - Examples: Addition Instruction, Debit-Credit, Kernels
  - Valid if one service is much more frequent than others
- Frequency of request types
  - Examples: Instruction mixes
  - Context sensitivity) Use set of services
  - History-sensitive mechanisms (caching)) Context sensitivity
- Time-stamped sequence of requests
  - May be too detailed
  - Not convenient for analytical modeling
  - May require exact reproduction of component behavior

Level of Detail (Cont)

- Average resource demand
  - Used for analytical modeling
  - Grouped similar services in classes
- Distribution of resource demands
  - Used if variance is large
  - Used if the distribution impacts the performance
- Workload used in simulation and analytical modeling:
  - Non executable: Used in analytical/simulation modeling
  - Executable workload: can be executed directly on a system

Representativeness

The test workload and real workload should have the same:

- Elapsed Time
- Resource Demands
- Resource Usage Profile: Sequence and the amounts in which different resources are used.

Timeliness

- Users are a moving target.
- New systems ⇒ new workloads
- Users tend to optimize the demand.
- Fast multiplication ⇒ Higher frequency of multiplication instructions.
- Important to monitor user behavior on an ongoing basis.

Other Considerations in Workload Selection

- Loading Level: A workload may exercise a system to its:
  - Full capacity (best case)
  - Beyond its capacity (worst case)
  - At the load level observed in real workload (typical case).
  - For procurement purposes ⇒ Typical
  - For design ⇒ best to worst, all cases
- Impact of External Components:
  - Do not use a workload that makes external component a bottleneck ⇒ All alternatives in the system give equally good performance.
- Repeatability

Summary

- Services exercised determine the workload
- Level of detail of the workload should match that of the model being used
- Workload should be representative of the real systems usage in recent past
- Loading level, impact of external components, and repeatability or other criteria in workload selection